Published online 2 December 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D237–D241***
doi:10.1093/nar/gkr1133

CharProtDB: a database of experimentally characterized protein annotations

Ramana Madupu*, Alexander Richter, Robert J. Dodson, Lauren Brinkac, Derek Harkins, Scott Durkin, Susmita Shrivastava, Granger Sutton and Daniel Haft

J Craig Venter institute, 9704 Medical Center Drive Rockville, MD 20850, USA

Received September 16, 2011; Revised October 31, 2011; Accepted November 8, 2011

ABSTRACT

CharProtDB (http://www.jcvi.org/charprotdb/) is a curated database of biochemically characterized proteins. It provides a source of direct rather than transitive assignments of function, designed to support automated annotation pipelines. The initial data set in CharProtDB was collected through manual literature curation over the years by analysts at the J. Craig Venter Institute (JCVI) [formerly The Institute of Genomic Research (TIGR)] as part of their prokaryotic genome sequencing projects. The CharProtDB has been expanded by import of selected records from publicly available protein collections whose biocuration indicated direct rather than homology-based assignment of function. Annotations in CharProtDB include gene name, symbol and various controlled vocabulary terms, including Gene Ontology terms, Enzyme Commission number and TransportDB accession. Each annotation is referenced with the source; ideally a journal reference, or, if imported and lacking one, the original database source.

INTRODUCTION

The process of biocuration can create a set of high-confidence annotations for a protein, separately asserting molecular function, preferred nomenclature for protein name and for gene symbol and assignment to one or more biological processes. Each of these annotations may be exploited for different purposes, such as supporting machine annotation of newly sequenced genomes or decorating nodes in multiple sequence alignment-based phylogenetic trees for protein functional inference (1). The advent of next-generation sequencing technologies and cheaper sequencing costs during the past decade has paved the way for sequencing a vast variety of genomes; completed prokaryotic genomes now number in the low thousands. This new abundance places every characterized protein into (often implicit) protein families and sets the stage for comparative genomics studies. Protein family co-occurrence across multiple taxa (phylogenetic profiling), conserved gene neighborhoods and metabolic context derived from pathway reconstruction can provide extensive guidance for the tricky process of using one characterized protein to annotate another. Previous attempts to select blanket generalizations such as prediction of equivalent enzymatic function at 50% identity or greater are too strict for some protein families, and too permissive for others (2). It is likely that new generations of annotation tools will use comparative genomics, libraries of prebuilt protein clusters and improved statistical models to achieve more accurate machine annotation directly from characterized proteins than has been possible from reliance on legacy annotation sets of mixed but unknown provenance.

Protein functional annotations deposited in public databases often represent inference by greatest sequence similarity to a protein with an ostensibly informative name and themselves lack traceable origins. These protein sequences then become the fundamental source for further BLAST-based propagation of protein functional assignments. Multiple types of 'transitive annotation error' can occur during such propagation of putative function, including overly specific annotation (3), founder effects that obscure functional diversity in large families such as radical SAM (4), daisy-chain inference that passes through non-overlapping regions of a multidomain protein (5) and faults from successive rounds of reinterpretation of an original protein name. Protein functional inference through computation will benefit, in the future, from increasingly deep comparative genomics resources. Conserved gene neighborhoods, pathway reconstructions and hole filling, multiple sequence alignments and molecular phylogenetic trees, identification of orthologs and paralogs and other data-driven techniques will help

*To whom correspondence should be addressed. Tel: 301 795 7871; Fax: 301 294 3142; Email: rmadupu@jcvi.org

propagate information with improving reach and accuracy. The sparse resource of proteins whose functions are known from direct laboratory characterizations will continue to grow in importance.

Anticipating that next-generation annotation tools will need to track which sequences carry primary annotations and to compute confidences during propagation, we have created a database architecture for representing experimentally derived protein characterizations, in which the original source of individual annotation fields is included. Gene Ontology (GO) (6) terms for both molecular function and biological process are presented with both provenance and GO evidence codes (ev-codes) to facilitate their use in machine annotation. We have established two methods for populating CharProtDB— manually as a synergistic benefit of biocuration of prokaryotic genomes and by import from various publicly accessible resources after filtering, processing, validation and consolidation of GO term assignments.

HISTORY

CharProtDB arose as a consequence of needing high quality annotations for the prokaryotic annotation projects at J. Craig Venter Institute (JCVI). Initially, it was just a listing of characterized accessions with a standardized name but grew to include annotation types listed in the content section (below). Initially, the primary emphasis was on experimentally characterized proteins useful for annotating prokaryotic pathogens, with a special focus on characterizations relevant for *Escherichia coli*, *Burkholderia*, *Bacillus* and *Clostridium*. The use of CharProtDB within JCVI automated annotation pipelines necessitated importing additional protein data sets with experimental evidence codes especially from model organism databases.

DATABASE

The central unit of CharProtDB is the protein record. Each protein record in CharProtDB (Figure 1) must have an assigned organism (by taxon ID), and at least one public accession, protein name and GO annotation complete with an experimental evidence code and an associated reference. The protein may also have one or more gene symbols assigned. Additional synonymous accessions are added to proteins, either automatically as the proteins are entered, or manually by curators. These synonymous accessions, in the context of CharProtDB, are limited to public accessions with both identical sequence and taxon ID.

Annotating using the GO system is of importance for several reasons; the GO system captures defined concepts (the GO terms) with unique ids, which can be attached to specific genes and the three controlled vocabularies of the GO allow for the capture of much more annotation information than is traditionally captured in protein common names, including, for example, not just the function of the protein, but its location as well. GO evidence codes implemented in CHAR directly correlate with the GO consortium definitions of experimental evidence codes (6).

Beyond GO annotations, the protein may be assigned one or more controlled vocabulary terms for enzyme functional classification, as Enzyme Commission (EC) numbers (http://www.chem.qmul.ac.uk/iubmb/) (7), or transporter functional classification, as Transport Classification (TC) numbers (8). Except for GO assignments, which must have a reference, any or all of these annotations may be linked to a reference. If a record was imported from an external database, the annotations coming from that database will be referenced back to the original source. Any additional references found, including those not directly linked to an annotation, will be attached to the applicable protein(s).

For leveraging CharProtDB in automated annotation, one protein name, one gene symbol and one or more GO terms, EC or TC annotations are marked as 'primary'. This represents the preferred choice for assignment to a predicted gene being automatically annotated. Apart from the primary annotation, CharProtDB also stores alternate protein names or synonyms and alternate gene symbol.

CONTENT

Data sources

The core of CharProtDB is a collection of prokaryotic proteins manually curated at JCVI. To that, we have added entries from the following databases that show explicit reference to a physical characterization: UniProtKB (9), EcoCyc (10), TCDB (Transporters) (8), MGOS (*Magnaporthe oryzae*) (11), AspGD (Aspergillus) (12), CGD (*Candida albicans*) (13) and GeneDB (*Schizosaccharomyces pombe*) (14).

To each of these, we have added characterized data from the GO Associations database. These entries have been flagged as being either fully characterized, or characterized for only one of the base GO assignments: process, function or component. At a lower level of confidence, we have added records from the above databases that have been marked as curated, but do not have biochemical characterizations (Table 1). We have developed an extensive list of controlled vocabulary terms that indicate the level characterization for a protein record and the source database from which it has been imported. A complete list of such 'status' terms is described in Table 1. We have begun adding records for proteins that may not have been functionally characterized through biochemical experiments, but only structurally through crystallography or proteomics or whose functional assertions have been made through bioinformatics analysis (15).

CharProtDB tools can link characterization data from multiple input streams through synonymous accessions or direct sequence identity. CharProtDB can represent multiple characterizations of the same protein, with proper attribution and links to database sources. As of publication, CharProtDB contains 16 046 proteins from 1588 species; 9185 proteins are bacterial in origin, with about one-third from each of *Enterobacteriales* and

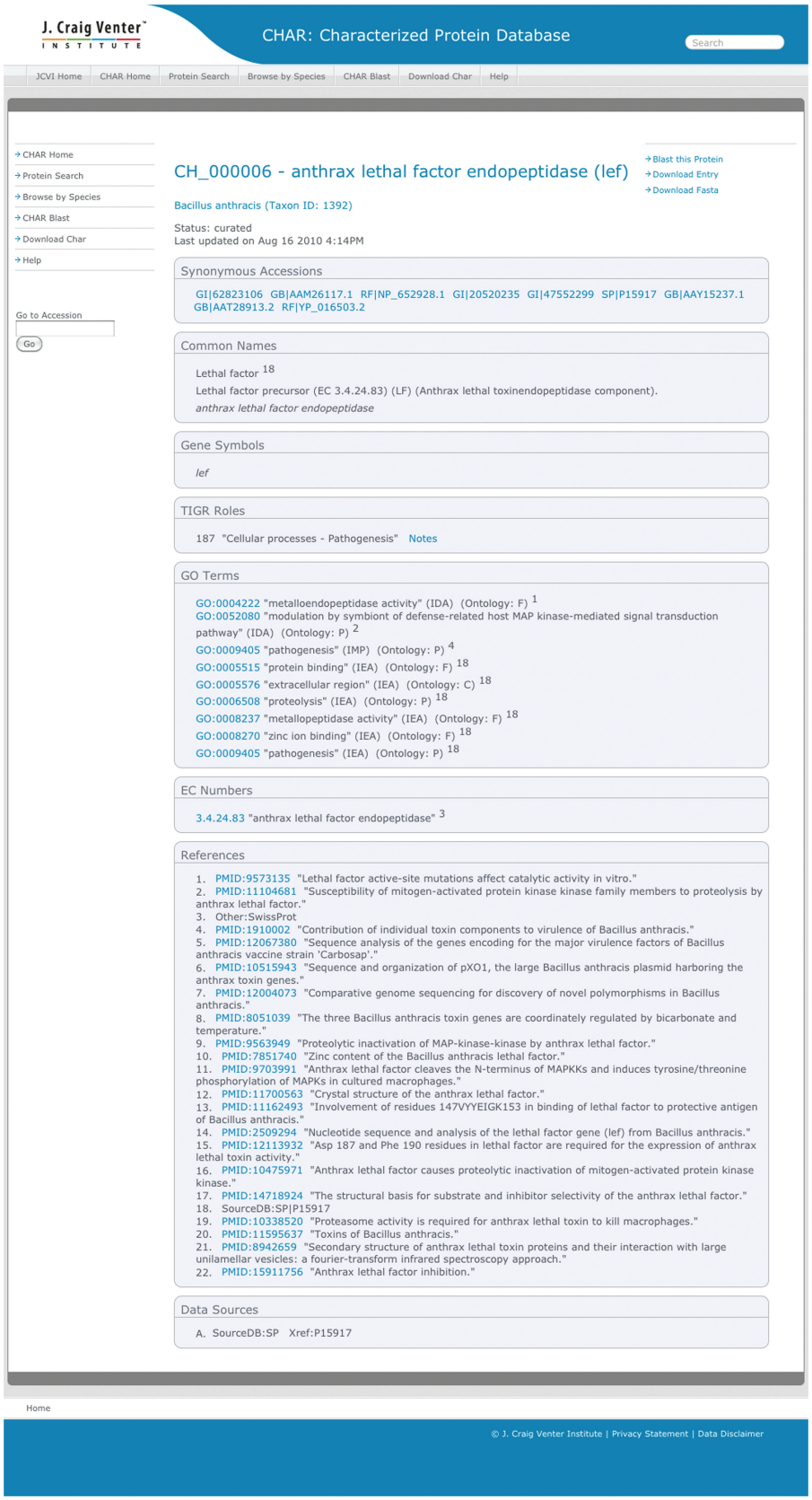

Figure 1. Detailed view of individual protein record.

Table 1. CharProtDB protein assertions

Status	Description	Proteins
curated	Proteins manually annotated by a JCVI annotator that contain both function and process annotation.	1075
curated_function	Proteins with only functional annotation, added by a JCVI annotator.	297
curated_process	Proteins with only biological process annotation, added by a JCVI annotator.	339
curated_component	Proteins with only cellular component annotation, added by a JCVI annotator.	7
curated_structure	Proteins with only structural annotation (e.g. proteomics or crystallographic data), added by a JCVI annotator.	39
curated_source	Proteins from a 'source' database marked as experimentally validated with added Gene Ontology annotation data.	6183
trusted_source	Proteins from a 'source' database marked as curated but without fully traceable experimental validation with added Gene Ontology annotation data.	8396

Bacillales; 5238 are eukaryotic in origin, with over three-fourths of those being fungal proteins; 931 proteins are viral in origin, primarily bacteriophage. Only 622 are archaeal in origin and are almost entirely imported from Swiss-Prot. Because of CharProtDB's origins and use as an internal resource for annotation, the species breakdown strongly reflects projects done at JCVI.

Database access and interface

CharProtDB is a standalone database that supports searching and retrieval of data using different search terms. A web interface allows users to search by protein name, protein accession, GO term, GO evidence code, gene symbol, EC number, organism name (genus or species), PubMed identifier or a combination of search terms. The complete list of protein records in CharProtDB broken down by taxonomic groups can be viewed on the website.

Data validation

Data imported into CharProtDB is extensively cross-validated, verified and standardized. We have developed several automated data consistency checks to resolve problems related to data discrepancy.

BLAST

Users can search against CharProtDB using a Blast utility. A BLAST sequence similarity search has been provided from the CharProtDB web interface, which accepts user input and can search the user submitted query sequence against the entire CharProtDB data set. Likewise the BLAST search utility available from a single protein view page provides convenient search capability for a single sequence search in CharProtDB against the entire database. The BLAST results are provided in standard BLAST text output, with links in the summary to the alignments and back to the protein details in CharProtDB.

Use of CharProtDB in automated annotation

AutoAnnotate is JCVI's automated prokaryotic functional annotation program designed for performing high throughput annotation of complete and draft bacterial genome sequences (16). Designed to assign 'heuristic annotation' controlled by parameters within the pipeline, the program weighs evidence from a ranked list of evidence types, annotates proteins according to molecular function and biological process, attaching both controlled vocabulary terms, such as GO terms supported by their appropriate GO evidence codes, and more human-readable fields such as the gene/protein name, gene symbol and EC number. AutoAnnotate primarily uses homology-based methods for automatic annotation. Homology evidence to CharProtDB proteins is given highest precedence in the ranking order. AutoAnnotate is the primary functional annotation pipeline adapted by the genome centers on the Human Microbiome project (HMP) to generate automated annotation of reference genomes (17). We have distributed the CharProtDB data set as part of JCVI's annotation pipeline to all the participating centers.

Access

The CharProtDB website can be accessed at (http://www.jcvi.org/charprotdb/). The CharProtDB is currently available freely for download as Swiss-Prot format records with all annotations, or just the sequences in Fasta format. Users can choose to download any displayed record, or the entire data set.

DISCUSSION AND CONCLUSIONS

CharProtDB is similar in goals to several other biocuration efforts that aim to provide computational access to assertions about experimentally verified protein function. NeXtProt (this issue), a resource for human proteins, is an example of an organism-specific database. Improvements to UniProtKB improve access by query to proteins with experimental evidence. COMBREX has begun an effort to enlist community annotators to contribute 'gold standard' biocuration of experimentally characterized proteins (18). Unfortunately, much work remains to be done to link experimental characterizations of protein function as reported in the literature with computationally accessible protein sequences, and much of the content of CharProtDB is unique. CharProtDB entries bring together consolidated protein annotations including sequence, synonymous accessions, GO annotations for

experimentally characterized proteins curated from scientific literature, a resource we found essential to enable best practices in microbial annotation. The CharProtDB proteins are available to the public as a source of computable objects, BLAST-ready and freely distributable protein set supported by querying interfaces. Although the set of 'trusted' category proteins obtained from external resources do not necessarily have direct experimental validation of function, they expand the collection of validated, certified entries in CharProtDB that can be used to annotate other proteins in a reliable way by automated annotation pipelines. The 'trusted' set can be filtered easily from the main curated data set using specific queries and a prefiltered set with only curated entries is provided for separate download.

ACKNOWLEDGEMENTS

The authors would like to thank past and present colleagues at the JCVI Bioinformatics and Information Technology departments for scientific contributions and technical support including Peter Rosanelli and Su Qi.

FUNDING

National Human Genome Research Institute (NHGRI) (R01 HG004881); National Institute of Allergy and Infectious Disease (contract HHSN266200100038C). Funding for open access charge: NHGRI.

Conflict of interest statement. None declared.

REFERENCES

1. Engelhardt,B.E., Jordan,M.I., Muratore,K.E. and Brenner,S.E. (2005) Protein molecular function prediction by Bayesian phylogenomics. *PLoS Comput. Biol.*, **1**, e45.
2. Rost,B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, **26**, 312–318.
3. Louie,B., Higdon,R. and Kolker,E. (2009) A statistical model of protein sequence similarity and function similarity reveals overly-specific function predictions. *PLoS One*, **4**, e7546.
4. Haft,D.H. and Basu,M.K. (2011) Biological systems discovery in silico: radical *S*-adenosylmethionine protein families and their target peptides for posttranslational modification. *J. Bacteriol.*, **193**, 2745–2755.
5. Galperin,M.Y. and Koonin,E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol.*, **1**, 55–67.
6. Gene Ontology Consortium. (2010) The Gene Ontology in 2010: extensions and refinements. *Nucleic Acids Res.*, **38**, D331–D335.
7. McDonald,A.G., Boyce,S. and Tipton,K.F. (2009) ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Res.*, **37**, D593–D597.
8. Saier,M.H. Jr, Noto,K., Tamang,D.G. and Elkan,C. (2009) The Transporter Classification Database: recent advances. *Nucl. Acids Res.*, **37**, D274–D278.
9. The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
10. Keseler,I.M., Santos-Zavaleta,A., Peralta-Gil,M., Gama-Castro,S., Muñiz-Rascado,L., Bonavides-Martinez,C., Paley,S., Krummenacker,M., Altman,T., Kaipa,P. *et al.* (2011) EcoCyc: a comprehensive database of *Escherichia coli* biology. *Nucleic Acids Res.*, **39**, D583–D590.
11. Soderlund,C.H.K., Pampanwar,V., Ebbole,D., Farman,M., Orbach,M.J., Wang,G.L., Wing,R., Xu,J.R., Brown,D., Mitchell,T. *et al.* (2006) MGOS: a resource for studying *Magnaporthe grisea* and *Oryza sativa* interactions. *Mol. Plant Microb. Interact.*, **19**, 1055–1061.
12. Arnaud,M.B., Costanzo,M.C., Crabtree,J., Inglis,D.O., Lotia,A., Orvis,J., Shah,P., Skrzypek,M.S., Binkley,G., Miyasato,S.R. *et al.* (2010) The Aspergillus Genome Database, a curated comparative genomics resource for gene, protein and sequence information for the Aspergillus research community. *Nucleic Acids Res.*, **38**, D420–D427.
13. Skrzypek,M.S., Costanzo,M.C., Inglis,D.O., Shah,P., Binkley,G., Miyasato,S.R. and Sherlock,G. (2010) New tools at the Candida Genome Database: biochemical pathways and full-text literature search. *Nucleic Acids Res.*, **38**, D428–D432.
14. Aslett,M. (2006) Gene Ontology annotation status of the fission yeast genome: preliminary coverage approaches 100%. *Yeast*, **23**, 913–919.
15. Selengut,J.D., Davidsen,T., Ganapathy,A., Gwinn-Giglio,M., Nelson,W.C., Richter,A.R. and White,O. (2007) TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res.*, **35**, D260–D264.
16. Davidsen,T., Ganapathy,A., Montgomery,R., Zafar,N., Yang,Q., Madupu,R., Goetz,P., Galinsky,K., White,O. and Sutton,G. (2010) The comprehensive microbial resource. *Nucleic Acids Res.*, **38**, D340–D345.
17. Nelson,K.E., Weinstock,G.M., Highlander,S.K., Worley,K.C., Creasy,H.H., Wortman,J.R., Rusch,D.B., Mitreva,M., Sodergren,E., Chinwalla,A.T. *et al.* (2010) A catalog of reference genomes from the human microbiome. *Science*, **328**, 994–999.
18. Roberts,R.J., Chang,Y.C., Hu,Z., Rachlin,J.N., Anton,B.P., Pokrzywa,R.M., Choi,H.P., Faller,L.L., Guleria,J., Housman,G. *et al.* (2011) COMBREX: a project to accelerate the functional annotation of prokaryotic genomes. *Nucleic Acids Res.*, **39**, D11–D14.